

**UNIVERSIDAD AUTÓNOMA DE MADRID
ESCUELA POLITÉCNICA SUPERIOR**



Grado en Ingeniería Informática

TRABAJO FIN DE GRADO

**Modelos de regresión local para predicción de
energía eólica**

Autor: María Barroso Honrubia

Tutor: Ángela Fernandez Pascual

abril 2021

Todos los derechos reservados.

Queda prohibida, salvo excepción prevista en la Ley, cualquier forma de reproducción, distribución comunicación pública y transformación de esta obra sin contar con la autorización de los titulares de la propiedad intelectual.

La infracción de los derechos mencionados puede ser constitutiva de delito contra la propiedad intelectual (*arts. 270 y sgts. del Código Penal*).

DERECHOS RESERVADOS

© 16 de abril de 2021 por UNIVERSIDAD AUTÓNOMA DE MADRID
Francisco Tomás y Valiente, n.º 1
Madrid, 28049
Spain

María Barroso Honrubia

Modelos de regresión local para predicción de energía eólica

María Barroso Honrubia

C\ Francisco Tomás y Valiente N.º 11

IMPRESO EN ESPAÑA – PRINTED IN SPAIN

AGRADECIMIENTOS

En primer lugar, quiero agradecer a la Cátedra UAM-IIC de Ciencia de Datos y Aprendizaje Automático la cesión de los datos de predicciones meteorológicas necesarios para los experimentos realizados en el proyecto, así como al Centro de Computación Científica (CCC) de la UAM por sus prestaciones computacionales.

Además, quiero dar las gracias a mi tutora, Ángela Fernandez, por compartir sus conocimientos conmigo y guiarme en estos últimos meses de mi carrera universitaria. También, agradecer a mis compañeros, y ya, amigos, cada día en la facultad, cada intensivo de estudio y cada brindis; me siento muy afortunada. Y por supuesto, no me olvido de todos los cafés de Diego y Lorenzo, que han sido necesarios para llegar hasta aquí.

Por último, agradecer a mi padre y a mi hermano el constante apoyo durante estos años. A mi abuela, por prepararme los tupperes de comida con tanto amor y, en especial, a mi madre, porque sé que estaría muy orgullosa de mí y porque la llevo siempre presente.

RESUMEN

El continuo desarrollo de las energías renovables, y en concreto, de la energía eólica, ha hecho necesario el uso de herramientas de predicción que permitan conocer con antelación la cantidad de energía eólica que será inyectada en la red. En este trabajo se propone un modelo de regresión local para predicción de energía eólica que define un modelo particular para cada punto del conjunto de prueba. Con este enfoque, se consiguen modelos precisos que no están dominados por muestras correspondientes a módulos de viento bajos y suponen una mejora en términos computacionales, sobre todo si el modelo es lineal.

Los modelos individuales contienen la información más parecida al punto a predecir al seleccionar los K vecinos próximos del conjunto completo de entrenamiento. En concreto, se utiliza la distancia euclídea para medir la similitud entre los vectores meteorológicos y cada modelo local se realiza utilizando el algoritmo de regresión lineal RR y el algoritmo de regresión no lineal SVR mediante un núcleo RBF. En el trabajo, se estudia el número K de puntos necesarios para entrenar un buen modelo local y, además, se comparan los resultados obtenidos tanto en términos computacionales como en términos de error con los resultados de los modelos globales RR y SVR. Además, en el caso de los modelos lineales, se examinan las relaciones inferidas para algunas características relevantes en predicción de energía eólica, como puede ser la velocidad del viento.

PALABRAS CLAVE

Energía eólica, Modelos locales, Ridge Regression, Support Vector Regression

ABSTRACT

The continuous development of renewable energies, and specifically, of wind energy, has made necessary the use of prediction tools that allow to know in advance the amount of wind power that will be injected into the power grid. In this work, we propose a local regression model for wind power forecasting that defines a particular model for each point of the test set. With this approach, and especially in linear methods, we achieve accurate models that are not dominated by low wind samples, and that implies an improvement also in computational terms.

The individual models contain the closest information to the point to be predicted by selecting the K nearest neighbors from the full training set. In particular, the Euclidean distance is used to measure the similarity between the meteorological vectors, and each local model is performed using the linear regression algorithm RR and the nonlinear regression algorithm SVR, with an RBF kernel. In the work, we study the number K of points needed to train a good local model and we also compare the results obtained, both in computational and error terms, with the results of the global RR and SVR models. Furthermore, in the case of the linear models, we examine the inferred relationships for some relevant characteristics to wind power forecasting, such as wind speed.

KEYWORDS

Wind power, Local models, Ridge Regression, Support Vector Regression

ÍNDICE

1	Introducción	1
1.1	Motivación de proyecto	1
1.2	Objetivos	2
1.3	Estructura del documento	2
2	Estado del arte en predicción de energía eólica	5
2.1	Métodos de predicción	5
2.2	Modelos meteorológicos	6
2.3	Evaluación de los resultados	7
3	Método propuesto para la predicción de energía eólica	9
3.1	Ridge Regression	9
3.2	Suport Vector Regression	10
3.3	Regresión local	12
3.4	Modelo de regresión local propuesto	12
4	Resultados experimentales	15
4.1	Herramientas usadas	15
4.2	Datos utilizados	15
4.2.1	Selección de coordenadas	16
4.2.2	División de datos	17
4.2.3	Estandarizado de datos	17
4.3	Experimento 1	18
4.3.1	Validación	18
4.3.2	Evaluación global	18
4.3.3	Selección de vecinos	18
4.3.4	Análisis de resultados	19
4.3.5	Análisis de modelos relevantes	19
4.4	Experimento 2	29
4.4.1	Validación	29
4.4.2	Evaluación	30
4.4.3	Análisis de resultados	30
4.5	Tiempo de cómputo	32
5	Conclusiones y trabajo futuro	35

Bibliografía	38
Apéndices	39
A Estadísticas descriptivas	41
A.1 Estadísticas modelo 31/12/2018 - 20:00:00	41
A.2 Estadísticas modelo 09/01/2018 - 07:00:00	41
A.3 Estadísticas modelo 13/02/2018 - 20:00:00	41
A.4 Estadísticas modelo 11/03/2018 - 14:00:00	45
A.5 Estadísticas modelo 15/12/2018 - 07:00:00	45

LISTAS

Lista de algoritmos

3.1 Algoritmo del modelo de regresión local propuesto	14
---	----

Lista de ecuaciones

Lista de figuras

1.1 Evolución de la energía eólica en España	1
3.1 Función de pérdida lineal	11
3.2 Ajuste modelo de regresión local	13
4.1 Coordenadas de la región de Galicia y alrededores	16
4.2 Producciones y predicciones locales y globales con Ridge Regression.	20
4.3 Gráfico de residuos del modelo local.	21
4.4 Matriz de correlación Ridge Regression global	22
4.5 Estadísticas para el modelo con fecha horaria 31/12/2018-20:00:00.	23
4.6 Matriz de correlación del modelo 09/01/2018-07:00:00.	24
4.7 Estadísticas para el modelo 09/01/2018-07:00:00.	25
4.8 Matriz de correlación para el modelo 13/02/2018-20:00:00.	26
4.9 Estadísticas para el modelo 13/02/2018-20:00:00.	27
4.10 Estadísticas para el modelo 11/03/2018-14:00:00.	28
4.11 Potencias horarias registradas del día 11/03/2018.	29
4.12 Predicciones horarias de SVR local y global.	31
4.13 Producciones y predicciones locales y globales con SVR.	31
A.1 Estadísticas modelo 31/12/2018 - 20:00:00	42
A.2 Estadísticas modelo 09/01/2018 - 07:00:00	43
A.3 Estadísticas modelo 13/02/2018 - 20:00:00	44
A.4 Estadísticas modelo 11/03/2018 - 14:00:00	46
A.5 Estadísticas modelo 15/12/2018 - 07:00:00	47

Lista de tablas

2.1	Horizontes de predicción	5
4.1	NMAE para Ridge Regression local seleccionando un número distinto de vecinos.	19
4.2	Residuos, predicciones y producciones de los modelos seleccionados.	21
4.3	Hiperparámetros de SVR y rejilla utilizada para encontrarlos.	30
4.4	NMAE para SVR local seleccionando distintos número de vecinos.	30
4.5	NMAE para SVR local validando cada modelo individual y variando K	32
4.6	Tiempo de cómputo de RR y SVR global.	32
4.7	Tiempo de cómputo de RR y SVR local.	33

INTRODUCCIÓN

1.1. Motivación de proyecto

La energía eólica lleva años experimentando un fuerte impulso, ya que, además de la reducción de emisiones contaminantes, cabe destacar la contribución que supone a la diversificación del balance energético, al progreso de la industria nacional, a la creación de nuevas empresas y a la consolidación de empleo en el entorno rural.

Una de las principales ventajas que ofrece y que ha impulsado en mayor medida su desarrollo respecto a otras fuentes de energía renovables es su elevada disponibilidad geográfica. Además, gracias a las mejoras en el comportamiento de los aerogeneradores, a la continua bajada en el coste por megavatio instalado de potencia eólica y a los avances en los métodos de apoyo a la programación y gestión, la capacidad total de energía eólica a nivel mundial ahora supera los 651 GW [1].

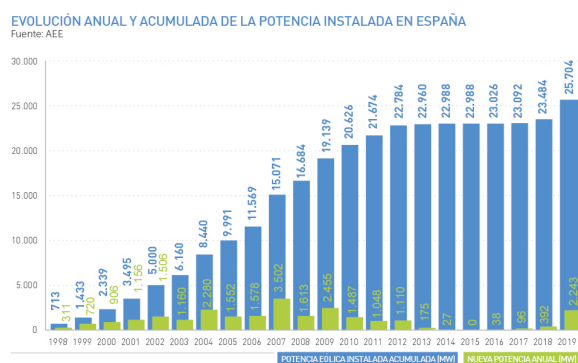


Figura 1.1: Evolución de la energía eólica en España [2]

En lo que respecta al caso español, al finalizar 2019, la energía eólica cubría el 21 % de la energía consumida y se alcanzaron los 26835 MW de potencia acumulada (Figura 1.1), con más de 1200 parques instalados en el territorio peninsular [2].

Además, dentro del marco regulatorio que impulsa las energías renovables en España mediante el Plan Nacional Integrado de Energía y Clima 2021-2030 [3], se han marcado como objetivos que

el 42 % del consumo de energía primaria en el año 2030 sea abastecido por energías renovables y que la producción eléctrica con fuentes renovables sea un 74 % del consumo bruto de electricidad. En concreto, se prevé una potencia instalada en el sector eléctrico de 157GW, de lo que 50GW serán energía eólica.

Todo este continuo desarrollo ha hecho necesario el uso de herramientas de predicción que permitan conocer con antelación la cantidad de energía eólica que será inyectada en la red y poder coordinar el resto de fuentes de generación, contribuyendo a minimizar tanto los costes de operación del sistema como a maximizar los beneficios o minimizar las penalizaciones de los agentes de mercado.

1.2. Objetivos

En este proyecto se ha marcado como primer objetivo explorar cómo se comporta un modelo cuando se selecciona, para su definición, sólo la información más parecida al punto a predecir. Para ello, se pretende definir un modelo diferente para cada punto del conjunto de prueba y estos modelos locales serán comparados contra modelos generales definidos sobre todo el conjunto de entrenamiento.

Debido a la necesidad de mejorar los modelos de predicción eólica, tanto en términos de error como en tiempo de cómputo, el segundo objetivo perseguido en el trabajo consiste en que el modelo local propuesto minimice el coste de cómputo sin que esto suponga una disminución en la precisión. Para comprobar su efectividad, se van a utilizar datos de energía eólica del parque de Sotavento, así como predicciones meteorológicas relacionadas con dicha energía. Se realizarán, para cada punto de prueba, modelos lineales y no lineales con los K vecinos próximos del conjunto completo de entrenamiento. Se estudiará el número de puntos K necesarios para entrenar un buen modelo y, en el caso de los modelos lineales, se examinarán las relaciones inferidas para algunas características relevantes en predicción la energía eólica, como puede ser la velocidad del viento.

1.3. Estructura del documento

Esta sección presenta de forma resumida la estructura y el contenido de cada una de las partes del trabajo.

En el Capítulo 1 se analiza el panorama actual en predicción de energía eólica, tanto a nivel mundial como a nivel nacional, y se fijan los objetivos del proyecto. En el Capítulo 2 se revisan los distintos enfoques y modelos existentes en la literatura para predicción de energía eólica. Se plantean además las dificultades que supone trabajar con variables meteorológicas y se presenta el método estándar de evaluación en predicción de energía. En el Capítulo 3 se definen en profundidad los dos algoritmos de regresión que serán utilizados en los experimentos: el algoritmo de regresión lineal Ridge Regres-

sion (RR) y el algoritmo de regresión no lineal Support Vector Regression (SVR). Una vez definidos, se introduce la idea general de los modelos de regresión local para finalmente presentar el modelo de regresión local propuesto en el trabajo. El Capítulo 4 describe los experimentos realizados y los resultados obtenidos para cada uno de ellos. Finalmente, el Capítulo 5 resume las conclusiones y aportaciones de este trabajo de investigación y se identifican posibles vías de mejora.

ESTADO DEL ARTE EN PREDICCIÓN DE ENERGÍA EÓLICA

En este capítulo se revisan los distintos enfoques y modelos existentes en la literatura para predicción de generación eólica. Se plantean las dificultades que supone trabajar con variables meteorológicas y se presenta un conocido método de evaluación en predicción de energía.

2.1. Métodos de predicción

La predicción de energía eólica puede clasificarse en función de los horizontes temporales o de la metodología aplicada [4]. Las 4 escalas de tiempo principales en las que pueden dividirse los horizontes temporales se resumen en la Tabla 2.1 y comprenden desde períodos de muy corto plazo, para el cual el rango de predicciones suele ser inferior a 30 minutos, hasta un rango de un mes en predicciones a largo plazo.

Horizonte de tiempo	Rango de tiempo	Aplicaciones
muy corto plazo	menos de 30 minutos	acciones de regulación, operaciones de la red en tiempo real, control de turbina
corto plazo	de 30 minutos a 6 horas	planificación y decisiones inteligentes de carga
medio plazo	de 6 horas a 1 día	seguridad operativa en el mercado eléctrico, comercio energético, decisiones de generación en línea y fuera de línea
largo plazo	de 1 día a 1 mes	requisitos de reserva, programas de mantenimiento, optimización del coste operativo, gestión de operaciones

Tabla 2.1: Horizontes de tiempo en predicción de energía eólica

Por otro lado, existen dos aproximaciones básicas para la predicción de la energía eólica, los modelos físicos y los modelos estadísticos.

Los modelos físicos utilizan caracterizaciones físicas [5] para modelar la velocidad del viento incidente en las turbinas de un parque. Posteriormente, se predice la potencia mediante la ecuación

$$p_e = \frac{1}{2} C_p \rho_{air} A_r v^3,$$

donde p_e es la potencia eléctrica generada, ρ_{air} es la densidad del aire, A_r es el área que atraviesa

el aire definida por el diámetro del rotor y v es la velocidad de la corriente de aire. C_p es el coeficiente de potencia, que indica el grado de aprovechamiento de la energía contenida en la corriente de aire.

Al contrario que en los modelos físicos, los modelos estadísticos no necesitan información del sistema, sino que se basan en relaciones lineales y no lineales entre los datos de predicciones meteorológicas (NWP) y la potencia generada. Este tipo de modelos se dividen a su vez en series temporales, modelos de aprendizaje automático y modelos híbridos, que son una combinación de estos.

Los modelos de predicción basados en series temporales fueron propuestos por Box-Jenkins, y tratan de extrapolar valores futuros que tomará una variable mediante el conocimiento y análisis de los valores pasados de dicha variable u otras variables explicativas. Algunos modelos de predicción de energía eólica basados en series temporales son ARMA [6], ARIMA [7] y ARMAX [8]. Este tipo de modelos necesitan poco tiempo de cálculo, son fáciles de formular y dan buenos resultados para predicciones a horizontes a corto plazo [9]. Sin embargo, su uso resulta poco útil para predicciones a horizontes superiores, hasta las 48 o 72 horas, que son las que suscitan más interés para los operadores del sistema y la participación en los mercados de energía.

Por su parte, los modelos de predicción basados en técnicas de aprendizaje automático ofrecen buenos resultados para este tipo de horizontes, siendo las redes neuronales artificiales (ANN) [10–12] y las máquinas de vectores de soporte de regresión [13] los algoritmos más comunes en predicción de energía eólica.

Además, en los modelos estadísticos para predicciones de energía eólica una buena opción es acudir al modelado local. Los modelos globales de predicción de energía eólica están dominados por muestras de velocidades de viento más bajas, y por tanto, el pronóstico de alta energía suele ser menos preciso. Se han realizado estudios de modelos locales basados en series temporales [14] utilizando ANN, Adaptive Neuro-Fuzzy Inference System (ANFIS), y Least Squares Support Vector Machine (LS-SVM), y se ha obtenido un rendimiento mejor que utilizando modelos globales. Lo mismo ocurre en [15], donde se construyen modelos específicos para predicción de alta energía al definir situaciones de alta energía en términos de los propios niveles de producción y utilizando un perceptrón multicapa (MLP).

En el presente trabajo se van a realizar predicciones a medio plazo utilizando técnicas de aprendizaje automático y acudiendo al modelado local.

2.2. Modelos meteorológicos

Como se ha comentado antes, para calcular predicciones de producción eólica es necesario disponer previamente de predicciones meteorológicas dadas en ubicaciones concretas ya que las variables meteorológicas son desconocidas a futuro. Además, se necesita que las características de entrenamiento y prueba sean las mismas, porque en caso contrario, el modelo no podría inferir relaciones

para predecir el resultado de potencia. Por tanto, se asume en el modelo un error intrínseco correspondiente a las predicciones meteorológicas.

En general, los pronósticos meteorológicos se basan en observaciones y modelos matemáticos que describen el comportamiento dinámico y físico de la atmósfera. Los modelos consisten en ecuaciones diferenciales parciales no lineales que proporcionan soluciones numéricas y presentan una gran dependencia de las condiciones iniciales para los que se definieron. De hecho, la forma en la que evoluciona la atmósfera es muy sensible a pequeños errores en el análisis inicial, de modo que cualquier ligera perturbación sobre las condiciones de partida puede derivar en importantes errores en poco tiempo [16].

Por tanto, los dos factores que influyen en la calidad de una predicción meteorológica son la incertidumbre en las condiciones iniciales de la atmósfera y las aproximaciones utilizadas por los modelos para representar los procesos físicos que ocurren en la misma [17]. Estos dos factores dan lugar a errores, que se amplifican al avanzar en el alcance temporal de la predicción y se propagan, a su vez, en los modelos de predicción de generación eólica.

Una de las agencias más importantes que proporciona predicciones numéricas sobre el tiempo atmosférico es el centro europeo de pronósticos meteorológicos a medio plazo (European Centre for Medium-Range Weather Forecasts, ECMWF). Los datos que produce se encuentran totalmente accesibles por los servicios meteorológicos de cada país integrante en la organización [18].

El modelo que utiliza se denomina Sistema de Predicción Meteorológica Integrado (IFS) y de manera general, modela las dinámicas de la atmósfera y los procesos físicos que ocurren en ella, como la formación de nubes, teniendo en cuenta que la atmósfera es caótica.

Las previsiones dadas por esta agencia serán las utilizadas en los experimentos del trabajo.

2.3. Evaluación de los resultados

El error de predicción de energía eólica se define como la cantidad de energía que se desvía de las observaciones reales de potencia en el intervalo de pronóstico de acuerdo a la siguiente ecuación

$$e(t+k|t) = W(t+k) - \hat{W}(t+k|t), \quad (2.1)$$

donde $W(t+k)$ es la potencia medida en el tiempo $t+k$ y $\hat{W}(t+k|t)$ es la potencia prevista para el tiempo $t+k$ calculada en el tiempo t . Para poder comparar el desempeño de modelos entre distintos parque eólicos con potencias instaladas diferentes, se normaliza el error con la potencia instalada en el parque de estudio en cada momento. De esta manera, el error normalizado se define como

$$e_N(t+k|t) = \frac{e(t+k|t)}{W_{nom}(t)}, \quad (2.2)$$

donde $W_{nom}(t)$ es la capacidad nominal del parque eólico en el tiempo t .

Existen diferentes métodos de evaluación propuestos en la literatura [19] para comparar modelos de predicción eólica en distintos casos de estudio. El error absoluto medio normalizado (NMAE) es una de las medidas de error más utilizadas para describir el error. Esto se debe a que, entre otras razones, es la métrica utilizada en los mercados eléctricos. Específicamente, esta medida está asociada al primer momento de la distribución de error de la predicción y es la media de los errores en valor absoluto de las predicciones calculadas respecto de las potencias reales

$$\text{NMAE}(k) = \frac{1}{N} \sum_{t=1}^N |e_N(t + k|t)|,$$

con e_N igual que en (2.1).

MÉTODO PROPUESTO PARA LA PREDICCIÓN DE ENERGÍA EÓLICA

En este capítulo se definen los dos algoritmos de regresión que serán utilizados durante los experimentos: el algoritmo de regresión lineal Ridge Regression y el algoritmo de regresión no lineal Support Vector Regression. Una vez definidos, se introduce la idea general de los modelos de regresión local para finalmente presentar el modelo de regresión local propuesto en el trabajo.

3.1. Ridge Regression

Ridge Regression [20] es un modelo de regresión lineal que trata de modelar la relación entre una variable continua y , denominada variable objetivo, y una o más variables independientes x_i , llamadas variables regresoras. La ecuación lineal a ajustar es

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \epsilon,$$

donde β_0 es la ordenada origen, β_j es el efecto promedio que tiene el incremento en una unidad de la variable regresora x_j sobre la variable dependiente y , manteniéndose constantes el resto de variables. Por su parte, ϵ es el residuo o error que queremos minimizar.

El método de ajuste se deriva del ajuste tradicional de mínimos cuadrados ‘encogiendo’ las estimaciones de los parámetros del modelo lineal mediante la penalización de sus valores. Dicha penalización tiene como consecuencia reducir la varianza del modelo, a costa de introducir un poco de sesgo, pero de forma que se optimicen ambas cantidades. De manera más específica, Ridge Regression busca minimizar esta otra función de coste

$$\sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij})^2 + \lambda \sum_{j=1}^p \beta_j^2 = RSS + \lambda \sum_{j=1}^p \beta_j^2,$$

donde λ es el *parámetro de regularización* y RSS es la suma de residuos al cuadrado.

El parámetro de ajuste λ controla el impacto o el nivel de regularización sobre los coeficientes. Cuando $\lambda = 0$, el término de penalización no tiene efecto, y el resultado es equivalente al de mínimos cuadrados. Cuando $\lambda \rightarrow \infty$, el impacto de esta penalización aumenta y el coeficiente de regresión

estimado se aproxima a 0.

En situaciones donde la relación entre la variable respuesta y los predictores es de tipo lineal o próxima a lineal, los estimadores de mínimos cuadrados tienen poco sesgo pero pueden presentar una varianza alta. Esto significa que un pequeño cambio en el conjunto de datos de entrenamiento puede hacer variar los coeficientes de manera sustancial. En particular, cuando p es casi igual de grande que n , los estimadores son extremadamente variables. Si $p > n$, los estimadores de mínimos cuadrados ni siquiera tienen un único valor y Ridge Regression reduce considerablemente la varianza a costa de un pequeño aumento en el sesgo (conforme λ aumenta). Por lo tanto, Ridge regression es una mejor opción en escenarios donde $p > n$.

3.2. Suport Vector Regression

La máquina de vectores de soporte de regresión [21] es un modelo de regresión que permite flexibilizar cuánto error es aceptable en el modelo.

Sea $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset \mathbb{X} \times \mathbb{R}$ el conjunto de datos de entrenamiento, donde \mathbb{X} denota el espacio de variables de entrada, por ejemplo, $\mathbb{X} = \mathbb{R}^p$. SVR tiene como objetivo encontrar una función $f(x)$ lo más simple posible que maximice la desviación de todas las variables objetivo muestreadas y_i hacia el valor del error o residuo ϵ . Estudiaremos el problema para el caso de funciones lineales ya que cualquier función no lineal puede convertirse en una función lineal mediante el uso de un núcleo.

Sea

$$f(x) = \langle w, x \rangle + b \text{ con } w, x \in \mathbb{X}, b \in \mathbb{R}.$$

Para garantizar la generalización del modelo, se penaliza la complejidad del mismo a través de la norma $\|w\|^2$. Para garantizar, además, que f se aproxime a todos los pares (x_i, y_i) con precisión ϵ , SVR se formula como un problema de optimización:

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 \\ & \text{sujeto a } |y_i - \langle w, x_i \rangle - b| \leq \epsilon. \end{aligned}$$

Este problema de optimización convexo no siempre es factible. Por tanto, es necesario permitir que algunos de los puntos y_i puedan encontrarse a una distancia de f mayor que ϵ . Para este propósito se introducen las variables de holgura ξ_i, ξ_i^* y la constante de regularización $C > 0$. Las variables de holgura miden la distancia de cada muestra al margen ϵ de f y C penaliza esta distancia para garantizar un equilibrio entre la complejidad de la función f y el ajuste a los datos. Así, el problema se reformula como:

$$\begin{aligned} & \min \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l (\xi_i + \xi_i^*) \\ \text{sujeto a } & \begin{cases} y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0. \end{cases} \end{aligned}$$

A medida que aumenta C , aumenta la tolerancia para los puntos fuera de ϵ . Cuando C se acerca a 0, la tolerancia se acerca a 0 y la ecuación es igual a la simplificada. Esto corresponde a tratar con la función de pérdida lineal ϵ -insensible definida por

$$|\xi|_\epsilon = \begin{cases} 0 & \text{si } \xi \leq \epsilon \\ |\xi| - \epsilon & \text{otro caso.} \end{cases} \quad (3.1)$$

Esta función de pérdida es simétrica y convexa. La convexidad es una condición necesaria para garantizar que el problema de optimización tenga una solución única que se pueda encontrar en un número finito de pasos. La Figura 3.1 muestra el problema lineal para una sola variable de entrada.

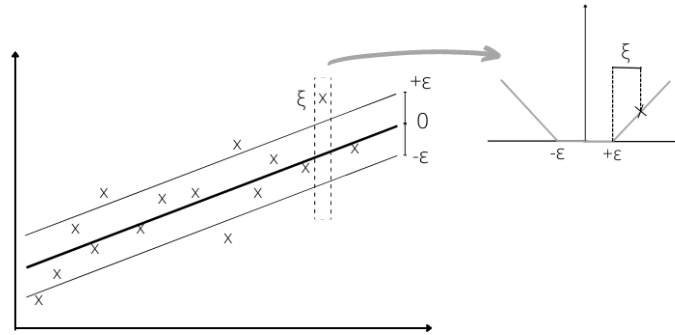


Figura 3.1: Configuración de la función de pérdida para una SVR lineal

A la hora de resolver el problema, lo formularemos como un problema dual para extender SVR a funciones no lineales. La solución dual para problemas lineales es de la forma

$$f(x) = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b, \quad (3.2)$$

con α_i y α_i^* variables duales. Es decir, la solución dual describe w como una combinación lineal de variables regresoras x_i . Aquellas variables x_i cuyos coeficientes no desaparecen reciben el nombre de *Vectores de Soporte*. Esto supone una de las principales ventajas de SVR: su complejidad computacional no depende de la dimensionalidad del espacio de entrada, sino del número de vectores de soporte.

Cuando la función no es lineal, los patrones de entrenamiento x_i se pueden mapear con $\phi : \mathbb{X} \rightarrow \mathcal{F}$ a un espacio \mathcal{F} dimensionalmente más alto, donde se pueda aplicar lo anterior. La solución al problema

no lineal queda determinada por

$$f(x) = \sum_{i=1}^l ((\alpha_i - \alpha_i^*)K(x_i, x)) + b,$$

con $K(x_i, x)$ la función específica del núcleo

$$K(x_i, x) = \langle \phi(x_i), \phi(x) \rangle.$$

Una de las funciones más conocidas y que será utilizada en el trabajo es la función gaussiana *RBF*, definida como

$$K(u, v) = \exp\{-\gamma\|u - v\|^2\},$$

donde $\gamma = \frac{1}{2\sigma}$, con σ la desviación estándar del conjunto de entrenamiento.

3.3. Regresión local

La regresión local, conocida como LOESS o LOWESS [22], es un enfoque de ajuste de curvas y superficies a datos mediante suavizados en los que el ajuste de una ecuación se realiza para cada dato de prueba utilizando únicamente observaciones en su entorno cercano. Una de sus principales ventajas es que elimina el ruido y permite observar tendencias y ciclos en los datos.

El método de estimación que resulta de este tipo de modelos consiste en definir, para cada dato, un entorno bajo alguna métrica. Dentro de ese entorno, se elige una regresión polinómica de grado bajo para modelar los datos y se estiman los parámetros del modelo que mejor se ajusten a las observaciones en el entorno elegido.

Para ilustrar la idea del modelo de regresión local se observa la Figura 3.2. Se busca modelar la relación existente entre una variable de entrada, que en este caso es la velocidad del viento a 100 metros en la coordenada más próxima a un parque eólico, y una variable de salida o variable respuesta, que es la potencia eólica generada. De la dispersión de puntos de la Figura 3.2, únicamente se ajustaría un modelo local a los puntos pertenecientes a los datos cuyo viento es similar al viento para el que se quiere calcular la predicción.

3.4. Modelo de regresión local propuesto

Una vez introducida la idea subyacente de un modelo de regresión local, se presenta el modelo de regresión local propuesto en el trabajo. El objetivo de este modelo es capturar la información local de un nuevo punto a predecir basándose en el conjunto de entrenamiento disponible. Para ello, el criterio

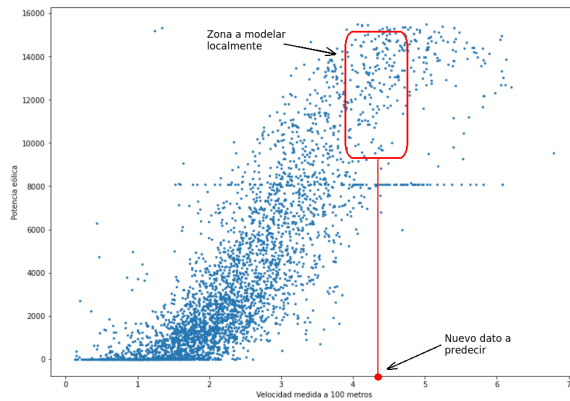


Figura 3.2: Ajuste de un modelo de regresión local a los datos cercanos. El eje x representa velocidades medidas a 100 metros en la región de un parque eólico, y el eje y representa la cantidad de energía eólica generada.

de selección del entorno de un dato x_i para el que quiere calcular la predicción, se realiza de acuerdo a sus K -vecinos más cercanos o K -NN. La medida de distancia utilizada para ello es la distancia euclídea, definida como

$$\sqrt{\frac{1}{p} \sum_{j=1}^p (x_i^j - x_j^j)^2},$$

donde x_i representa el dato a predecir y los puntos x_j representan el conjunto de datos de entrenamiento sobre los que ajustar la curva.

En un método de regresión local, el proceso de validación y el proceso de ajuste al entorno más cercano de un dato no se realiza hasta que queremos calcular la predicción para ese dato. Sin embargo, en el modelo propuesto, los parámetros utilizados para ajustar cada modelo local individual se eligen de manera general validando ese mismo modelo con el total de datos de entrenamiento. El método propuesto permite utilizar cualquier modelo de regresión como predictor. En el presente trabajo se van a utilizar, en concreto, los dos modelos de regresión paramétricos definidos al principio de este mismo capítulo: Ridge Regression y Support Vector Regression. Finalmente, la evaluación se realiza utilizando la medida NMAE, descrita en la Subsección 2.3.

La definición general del método se muestra en el Algoritmo 3.1, que describe la fase de validación o búsqueda de hiperparámetros globales, la estandarización de los datos de entrenamiento y prueba de acuerdo a la desviación típica de los datos de entrenamiento, la fase de entrenamiento local individual y por último, la evaluación del modelo.

```
input :  $\{X, Y\}$  – Dataset  
         $K$  - Number of nearest neighbors  
output:  $\{X_{test}, Y_{pred}\}$  – Prediction dataset  
1  $X_{train}, X_{test}, Y_{train}, Y_{test} \leftarrow \text{split}(X, Y);$   
2  $\text{standardScaler}(X_{train}, X_{test});$   
3  $\text{validation}(X_{train}, Y_{train});$   
4 for  $(x_{test}, y_{test})$  in  $(X_{test}, Y_{test})$  do  
5    $K_x, K_y \leftarrow \text{KNN}(X_{train}, Y_{train}).\text{kneighbors}(x_{test}, K);$   
6    $\text{fit}(K_x, K_y);$   
7    $Y_{pred} \leftarrow \text{predict}(x_{test});$   
8 end  
9  $\text{NMAE} \leftarrow \text{evaluation}(Y_{test}, Y_{pred});$ 
```

Algoritmo 3.1: Algoritmo de validación, entrenamiento y evaluación del modelo de regresión local propuesto.

RESULTADOS EXPERIMENTALES

Ya propuesto en el capítulo anterior el modelo de regresión local para predicción de energía eólica, procedemos a someterlo a diferentes experimentos para probar su eficacia y viabilidad. Se realizarán pruebas tanto para el modelo local basado en Ridge Regression, como para el modelo local basado en Support Vector Regression y se compararán los resultados de estos modelos locales con sus modelos globales correspondientes. Por último, se hará un breve estudio sobre la capacidad de cómputo requerida por cada uno de los modelos.

4.1. Herramientas usadas

Todos los experimentos se han desarrollado en Python 3.6.12, haciendo uso de la librería Scikit-Learn, numpy y pandas. Los cálculos que involucran los datos descritos a continuación, por ser de carácter privado, se ejecutaron en el Centro Computacional Científico de la UAM. Algunas de las características del servidor utilizado consisten en 40 cores, 700GB RAM, 40 TB disco duro y 2 GPUs tesla p100. Los cálculos se ejecutaron mediante el envío de trabajos a un sistema de colas, que es controlado por un gestor de recursos llamado Slurm. También se utilizó Jupyter Notebook para analizar resultados y generar gráficos con Matplotlib.

4.2. Datos utilizados

Para evaluar los modelos propuestos, en este trabajo se utilizarán datos de producción eólica del Parque Eólico Experimental Sotavento y datos de predicciones meteorológicas previstas por el Centro Europeo de Previsiones Meteorológicas a Plazo Medio [18]. Concretamente, se utilizan datos horarios registrados durante los años 2016, 2017 y 2018.

Los datos de producción eólica de Sotavento correspondientes a las potencias medias de cada hora, medidas en kW. Por su parte, los datos de predicciones meteorológicas vienen dadas en una rejilla de resolución 0.125° que cubre la península ibérica. Incluyen las siguientes variables:

- 10u: Velocidad de la componente u del viento medida a 10 metros (m/s).
- 10v: Velocidad de la componente v del viento medida a 10 metros (m/s).
- 100u: Velocidad de la componente u del viento medida a 100 metros (m/s).
- 100v: Velocidad de la componente v del viento medida a 100 metros (m/s).
- 2t: Temperatura medida a 2 metros (K).
- sp: Presión medida a nivel de superficie (Pa).
- 10Vel: Módulo de la velocidad del viento medida a 10 metros (m/s).
- 100Vel: Módulo de la velocidad del viento medida a 100 metros (m/s).

Dado que la velocidad es una magnitud vectorial, el vector velocidad horizontal del viento se divide en las componentes presentadas arriba: la componente zonal u y la componente meridional v. La componente zonal es la componente de la velocidad horizontal a lo largo de un círculo de latitud, en dirección Oeste a Este. Por su parte, v es la componente de la velocidad horizontal a lo largo de un meridiano, de Sur a Norte.

La necesidad de utilizar predicciones a 10 y a 100 metros, en lugar de utilizar las predicciones en superficie, se debe a que la altura de los aerogeneradores de Sotavento está entre los 40 y 60 metros. Por tanto, se requiere de ambas predicciones para obtener resultados más precisos.

Inicialmente, se disponía de datos horarios en 435 coordenadas seleccionadas alrededor de Galicia. Las coordenadas pueden visualizarse en la Figura 4.1.

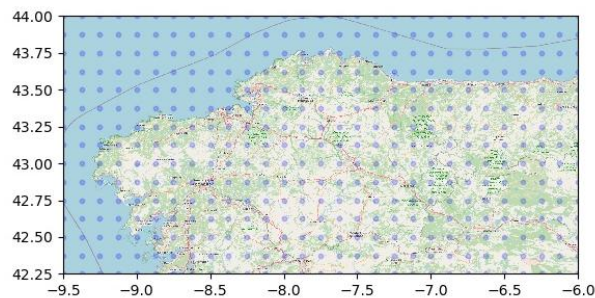


Figura 4.1: Coordenadas de la región de Galicia y alrededores para las que se han medido las predicciones meteorológicas. El mapa de la imagen ha sido extraído de <http://openstreetmap.org/> (febrero 2021).

4.2.1. Selección de coordenadas

Para evaluar nuestros modelos, nos interesan aquellas coordenadas cercanas a Sotavento. En concreto, sólo se utilizarán las variables meteorológicas previstas en las 9 coordenadas más cercanas al parque. Esto supone una reducción de la dimensionalidad del espacio de entrada de 3480 a 72 variables y, por tanto, un menor coste computacional.

El Parque Eólico Experimental Sotavento está situado en las coordenadas geográficas

N: 43.354377° W: -7.881213°,

por lo que al definir una rejilla de 9 puntos a su alrededor, seleccionamos los puntos

(43.5, -8)	(43.5, -7.875)	(43.5, -7.75)
(43.375, -8)	(43.375, -7.875)	(43.375, -7.75)
(43.25, -8)	(43.25, -7.875)	(43.25, -7.75).

4.2.2. División de datos

El proceso de entrenamiento o ajuste del modelo propuesto se realiza sobre los datos horarios registrados en 2016 y 2017, tomando un subconjunto que se corresponde al entorno más cercano del dato a predecir. Por su parte, el proceso de evaluación se realiza respecto a los datos tomados en 2018. Por tanto, las dimensiones de los conjuntos de entrenamiento y prueba son 17544 x 72 y 8760 x 72, respectivamente.

4.2.3. Estandarizado de datos

Los métodos de regularización propuestos, Ridge Regression y Support Vector Regression, actúan sobre la magnitud de los coeficientes del modelo local. Las variables regresoras deben estandarizarse utilizando la fórmula

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

para que todas estén en la misma escala. El denominador es la desviación estándar estimada del j -ésimo regresor. En consecuencia, todos los regresores estandarizados tendrán una desviación estándar de uno y el ajuste final no dependerá de la escala en la que se midieron.

La desviación estándar de los datos de entrenamiento es la estadística utilizada para estandarizar tanto el conjunto de entrenamiento como el conjunto de prueba. A partir de ahora, aunque no se haga referencia explícita a ello, asumimos que los datos que manejan los modelos son valores estandarizados.

4.3. Experimento 1

Una vez preprocesados los datos, se realizará un proceso de validación para obtener la constante que mejor ajusta nuestro modelo Ridge Regression sobre todos los datos de entrenamiento. A continuación, se evaluará el modelo local sobre el conjunto de prueba seleccionando distintos K vecinos más cercanos del conjunto de entrenamiento. Para el K que mejor precisión se obtenga, se hará un análisis exhaustivo de las características de los vecinos seleccionados para algunas predicciones relevantes.

4.3.1. Validación

El único parámetro en el modelo Ridge Regression es la constante λ de regularización. Se ha de encontrar el mejor parámetro que se ajuste a los datos de entrenamiento, y ese valor será utilizado en cada modelo local. Tras lanzar un proceso de validación cruzada de 5 particiones utilizando una búsqueda en rejilla o grid search sobre $\{10^k : -1 \leq k \leq 3\}$, el menor error obtenido corresponde a trabajar con $\lambda = 100$. Recordemos que un valor alto en el parámetro de regularización supone una reducción de la varianza a costa de un aumento de sesgo con el fin de que nuestro modelo generalice mejor.

4.3.2. Evaluación global

Evaluando un primer modelo global Ridge Regression con la constante λ anterior sobre los datos de prueba de 2018, obtenemos un NMAE igual a **8.44 %**. Este error está lejos del error obtenido en el estado del arte para predicciones de energía eólica en este parque [13] y el objetivo del experimento es mejorarlo utilizando el modelo local propuesto.

4.3.3. Selección de vecinos

Como se comentó en el capítulo anterior, el modelo local propuesto utiliza K -NN para seleccionar el subconjunto de datos de entrenamiento sobre los que ajustar el modelo. Los valores de K que se utilizan para evaluar el modelo local fueron elegidos redondeando distintos porcentajes sobre los 17544 datos totales de entrenamiento. La Tabla 4.1 muestra el porcentaje de error cometido, para cada K seleccionado, al evaluar las predicciones de los datos de 2018 tomando NMAE como indicador.

Viendo los resultados obtenidos de NMAE para cada K , podemos empezar a extraer ciertas conclusiones, como el hecho de que un mayor número de vecinos supone un mayor error en las predicciones. Además, este error se aproxima, tal y como esperábamos, al del modelo global. Este resultado tiene

%	0.01	0.025	0.05	1	5	10	25	50	75
K	17	44	87	175	877	1754	4386	8772	13158
NMAE (%)	6.92	6.83	6.88	6.94	7.15	7.26	7.53	7.95	8.22

Tabla 4.1: NMAE para Ridge Regression local seleccionando un número distinto de vecinos.

sentido ya que el parámetro λ utilizado es fijo. Por otro lado, el modelo local más preciso se obtiene al seleccionar un 0.025 % de datos de entrenamiento, es decir, 44 vecinos, cometiendo un error igual a 6.83 %. Las predicciones en Sotavento se consideran aceptables cuando este error es menor que 7 %, por tanto hemos conseguido que un modelo lineal como Rige Regression sea un modelo competitivo al hacerlo local.

En las siguientes subsecciones se presenta un análisis de los resultados obtenidos con el modelo de menor error, es decir, el que utiliza 44 vecinos.

4.3.4. Análisis de resultados

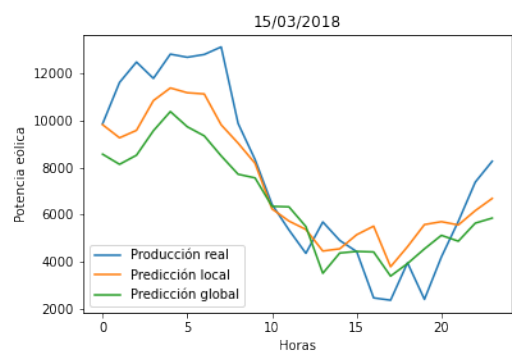
A continuación, se analizan y comparan las predicciones obtenidas por el modelo Ridge Regression local con 44 vecinos y las predicciones del modelo global frente a las producciones reales de diferentes días a lo largo del año 2018. Para ello, se han seleccionado días con producciones altas y bajas, así como situaciones con mucha y poca dispersión.

En las gráficas de la Figura 4.2 puede observarse la tendencia de las predicciones locales a situarse cerca de las producciones reales manteniendo la tendencia de la misma. Además, hay una notable mejora en la precisión del modelo local frente a la precisión del modelo global, sobre todo para potencias bajas.

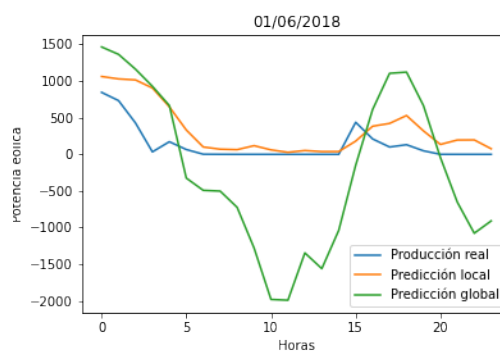
4.3.5. Análisis de modelos relevantes

Nos interesa también conocer cómo selecciona el modelo local los 44 datos de entrenamiento para definir los modelos individuales de cada punto de prueba. En concreto, estudiaremos las características de estos vecinos para algunos vectores de prueba que den tanto buenas como malas predicciones con Ridge Regression. Además, se conocerán las variables que aportan más información a estos modelos, es decir, aquellas que están más correlacionadas con la variable objetivo. Todo este análisis es posible ya que estamos trabajando con un modelo de regresión lineal que nos permite evaluar el efecto de cada predictor en presencia del resto, así como interpretar los coeficientes estimados.

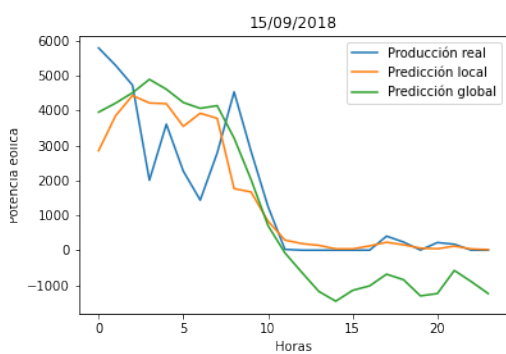
Para seleccionar aquellos modelos que nos ofrecen información interesante, prestaremos atención a los residuos. La Figura 4.3 muestra el gráfico de residuos obtenidos para las producciones de



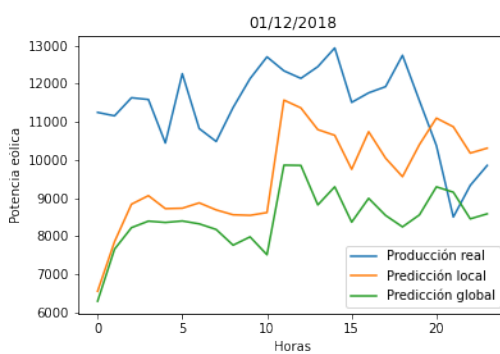
(a) 15 de Marzo de 2018



(b) 1 de Junio de 2018



(c) 15 de Septiembre de 2018



(d) 1 de Diciembre de 2018

Figura 4.2: Producciones y predicciones locales y globales con Ridge Regression en diferentes días a lo largo del año 2018.

Sotavento utilizando el modelo local con 44 vecinos. Los residuos han sido calculados utilizando la diferencia entre producciones y predicciones sin el valor absoluto para que los datos estén más dispersos y puedan interpretarse mejor.

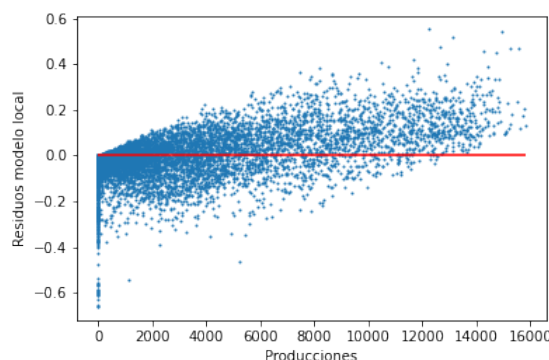


Figura 4.3: Gráfico de residuos del modelo local.

De la gráfica de residuos concluimos que para producciones altas, el método propuesto infraestima la predicción. Por el contrario, las producciones demasiado bajas son sobreestimadas por el modelo. Además, observando la gráfica también podemos deducir que para las producciones nulas registradas, los altos residuos se deben a horas en las que las máquinas pararon su producción por labores de mantenimiento. Estos residuos, sumados a aquellos derivados de errores meteorológicos, son errores con los que cuenta nuestro modelo y le resta precisión.

Para continuar con nuestro análisis, se seleccionan 4 modelos con residuos mínimos y máximos. Sus predicciones y producciones correspondientes se observan en la Tabla 4.2.

Residuos	Fecha horaria	Predicción local	Producción	Predicción global
0	31/12/2018 - 20:00:00	0	0	1.694
1.41e-06	09/01/2018 - 07:00:00	6563.40	6563.43	7470.49
0.55	13/02/2018 - 20:00:00	2573.31	12255.83	3153.94
0.659	11/03/2018 - 14:00:00	11582.38	0	10818.26

Tabla 4.2: Residuos, predicciones y producciones de los modelos seleccionados.

Previamente a analizar cada modelo seleccionado, haremos un análisis del modelo global. En concreto, nos fijamos en la matriz de correlación de la Figura 4.4 para evaluar las características que aportan más información a la función objetivo.

Si prestamos atención a la última fila de la matriz, se observa la alta correlación de la velocidad del viento sobre la potencia registrada. Esta correlación es elevada para la velocidad del viento tanto a 10 como a 100 metros en las 9 coordenadas alrededor de Sotavento. Por otro lado, podemos observar como la temperatura y la presión apenas son relevantes en el modelo global. También se aprecia la

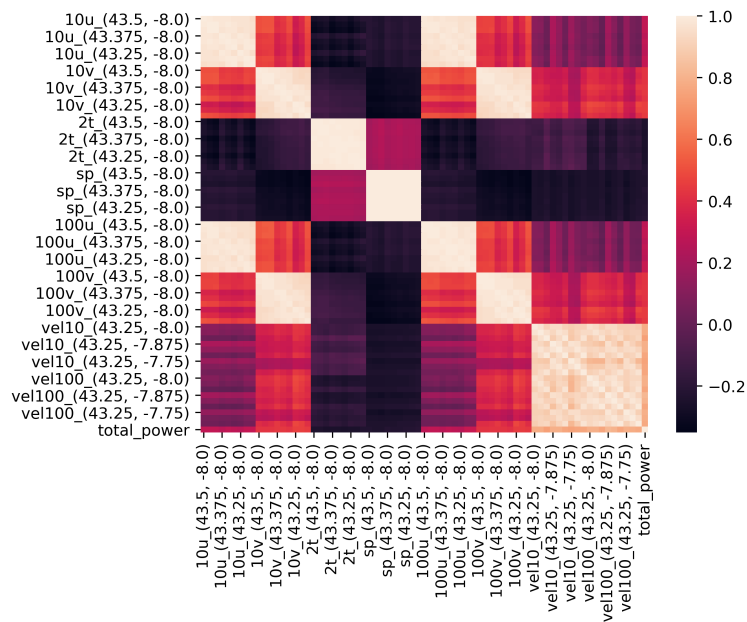


Figura 4.4: Matriz de correlación Ridge Regression global.

alta correlación que existe entre las componentes del viento a 10 y 100 metros.

A continuación, analizaremos los modelos individuales seleccionados. Este análisis se llevará a cabo mediante la visualización de algunas estadísticas descriptivas utilizando diagramas de caja BoxPlot. Este tipo de gráficos muestran un resumen de medidas descriptivas, como medianas y cuartiles, y nos permiten identificar valores atípicos y comparar las distribuciones. La información se interpreta de la siguiente manera:

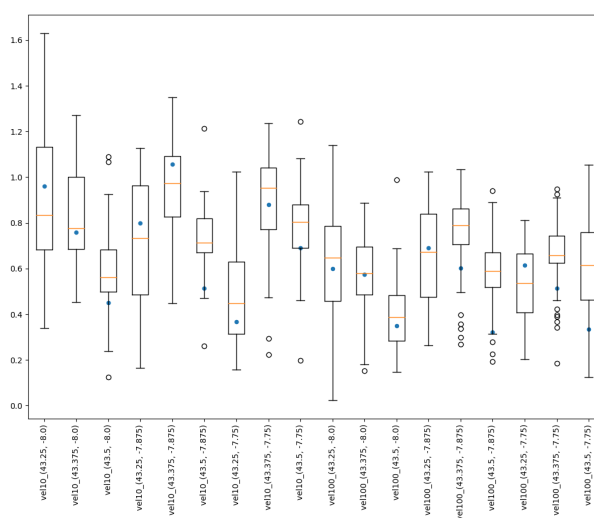
- Mediana. Valor central de la muestra. Se representa por la línea horizontal dentro de la caja.
- Primer Cuartil (Q_1). Valor de la variable tal que el 25 % de las observaciones se encuentran por debajo de este valor. Se presenta mediante el límite inferior de la caja.
- Tercer Cuartil (Q_3). Valor de la variable tal que el 75 % de las observaciones se encuentran por encima de este valor. Se presenta mediante el límite superior de la caja.
- Rango Inter cuartílico (RIC). Es la diferencia entre el tercer y el primer cuartil.
- Límites Superior o Inferior (Ls o Li). El valor mínimo Li de los datos está determinado por el valor de $Q_1 - 1,5(Q_3 - Q_1)$ mientras que el valor máximo de los datos está determinado por la fórmula $Q_3 + 1,5(Q_3 - Q_1)$. Se representan por el límite de las prolongaciones de la caja.
- Los valores atípicos son aquellos que están más allá de los límites inferior y superior.

En el trabajo sólo se mostrarán las estadísticas descriptivas correspondientes para algunas variables relevantes, que son, en la mayoría de los casos, la velocidad del viento y sus componentes. Aún así, el lector interesado puede encontrar en el Apéndice A el resto de estadísticas para todos los modelos seleccionados.

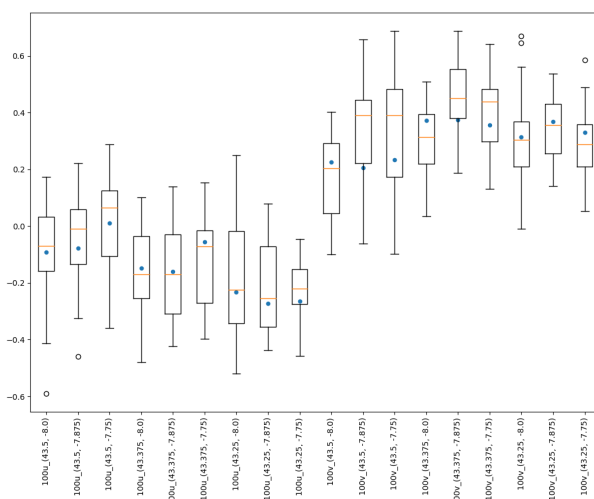
Modelo 31/12/2018-20:00:00

Este modelo predijo de manera exacta una potencia nula. Para este modelo los coeficientes estimados son todos igual a 0. Este es un caso curioso en el análisis de regresión lineal. Todos los datos seleccionados para ajustar el modelo tienen potencias iguales a 0 y no pueden calcularse las correlaciones entre las variables regresoras y la función objetivo. Aún así, veamos qué estadísticas presentan los datos seleccionados.

En la Figura 4.5 se visualizan estadísticas de las velocidades del viento a 10 y a 100 metros en las distintas coordenadas y sus componentes del viento u y v a 100 metros. Se han elegido estos datos por ser especialmente representativos para predicciones de potencias eólicas.



(a) Vel10 y Vel100



(b) 100u y 100v

Figura 4.5: Estadísticas de las velocidades a 10 y 100 metros 4.5(a), y las componentes u y v del viento a 100 metros 4.5(b) para el modelo con fecha horaria 31/12/2018-20:00:00.

Como podemos observar, las velocidades del viento tanto a 10 como a 100 metros para todos los datos seleccionados son menores que 2. Los aerogeneradores no generan energía con valores de velocidad tan bajos, por tanto, las potencias nulas observadas son coherentes. Respecto a las componentes del viento, vemos que los valores son de nuevo muy bajos. Además, la selección de vecinos cercanos se realizó con éxito ya que las características meteorológicas del dato a predecir en la mayoría de las coordenadas están dentro del rango intercuartílico y próximos a las medianas.

Modelo 09/01/2018-07:00:00

Este modelo predijo una potencia media de 6563.40 kW, 3 centésimas menos de lo esperado. De nuevo se trata de una buena predicción, y estudiaremos las estadísticas de algunas de las características de los vecinos seleccionados para ajustar el modelo.

Las variables más correlacionadas con la potencia total generada en Sotavento pueden apreciarse en la matriz de correlación de la Figura 4.6. Estos valores no son muy altos y se ha comprobado que son similares para todas las características en los diferentes modelos seleccionados. Esto se debe a que el número de características (72) es mayor al número de datos de entrenamiento (44).

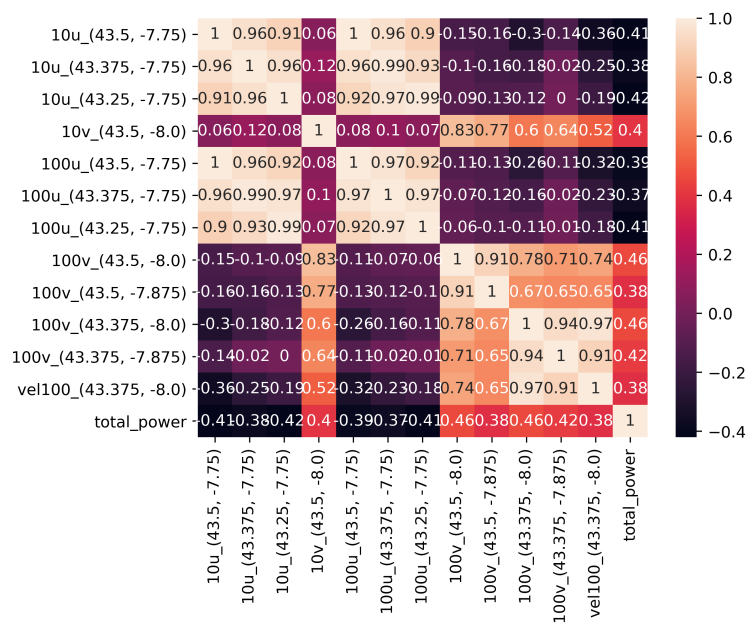
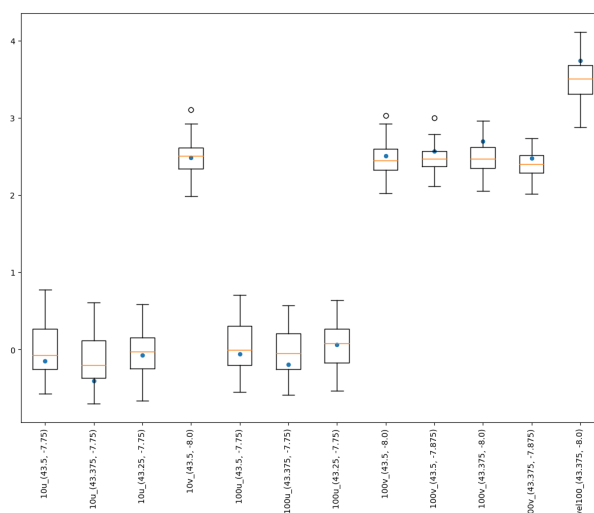


Figura 4.6: Matriz de correlación de las 12 variables más significativas del modelo 09/01/2018-07:00:00.

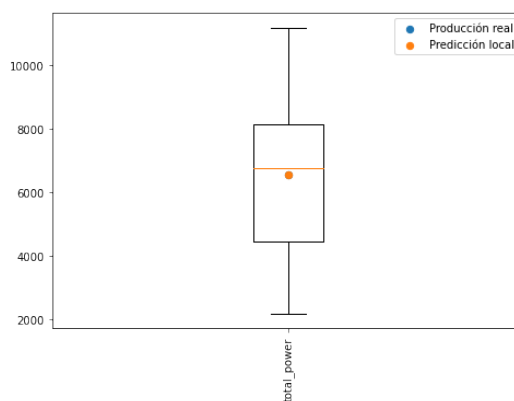
Aún así, se analizarán las estadísticas para estas 12 variables, presentes en la gráfica 4.7(a). Los vecinos seleccionados presentan componentes u del viento concentradas en cero y componentes v concentradas en torno a un valor de 2.5. Además, el dato a predecir se encuentra cerca de la mediana de estas características en la mayoría de las coordenadas elegidas. También pueden verse las estadísticas para la velocidad de viento en la coordenada más cercana a Sotavento, alcanzando velocidades

de 3.5 m/s.

Finalmente, si nos fijamos en las producciones registradas y esperadas de la gráfica 4.7(b), vemos que los vecinos seleccionados son aquellos con potencias registradas entre 4000 y 8000 kW y el dato a predecir se encuentra muy próximo a la mediana. No se muestra la producción real por tomar un valor cercano a la predicción local.



(a) 12 variables más significativas



(b) Potencia

Figura 4.7: Estadísticas de las 12 variables más significativas 4.7(a) y de la variable objetivo 4.7(b) para el modelo con fecha horaria 09/01/2018-07:00:00.

Modelo 13/02/2018-20:00:00

En este caso se predijo una potencia de 2573.31 kW frente a la potencia real de 12255.83 kW. La última fila de la matriz de correlación de la Figura 4.8 muestra la temperatura, la componente u del viento y las velocidades como las características más significativas para la variable objetivo.

En los diagramas 4.9(a) y 4.9(b) de la Figura 4.9 se observa una buena selección de vecinos

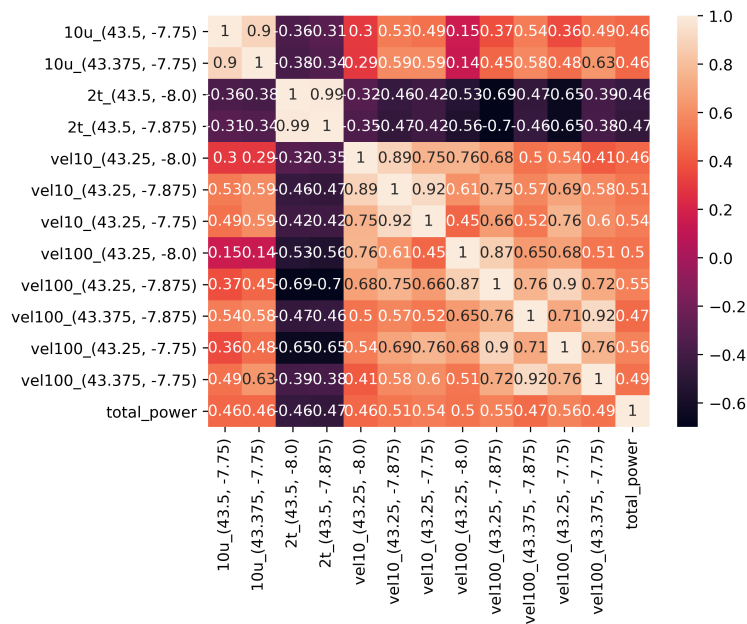


Figura 4.8: Matriz de correlación de las 12 variables más significativas para el modelo 13/02/2018-20:00:00.

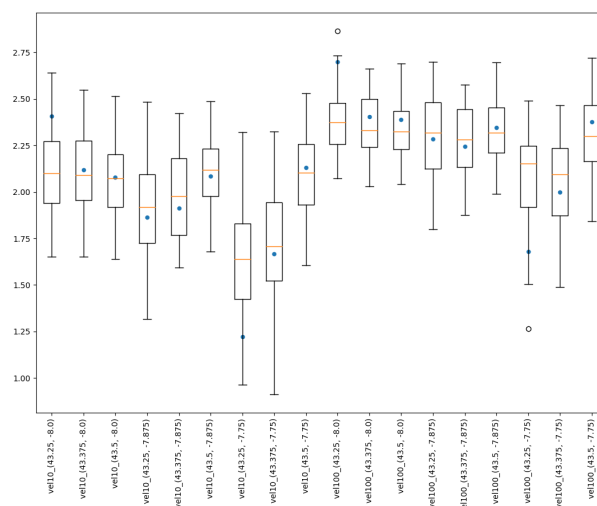
próximos. En la mayoría de las coordenadas el dato a predecir se encuentra o cerca de la mediana o en su rango intercuartílico. Además, la predicción de potencia coincide con la mediana de las potencias de los datos seleccionados. Sin embargo, el valor real de potencia generada, 12255.83 kW, es muy alto comparado con las potencias producidas en los puntos seleccionados, como se puede observar en la Gráfica 4.10(c).

En el Apéndice A.5, puede observarse un modelo que predijo con gran precisión una alta potencia, concretamente 12264 kW. Se ha seleccionado este modelo con la idea de comparar las estadísticas de los vecinos seleccionados por ambos modelos, ya que se esperaría que nuestro modelo hubiese hecho una selección similar a este. Sin embargo, todas las estadísticas descriptivas de las variables toman valores más altos que los que presenta nuestro dato a predecir.

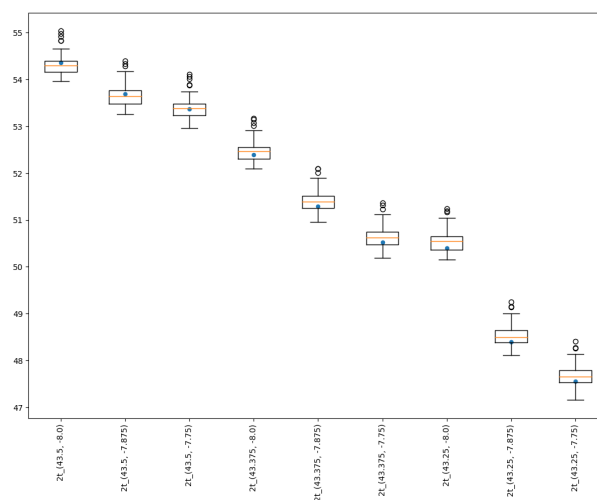
Este caso tan particular, donde parece que los puntos seleccionados se parecen al punto a predecir, pero donde las predicciones son tan dispares, puede deberse a un error en las predicciones meteorológicas para esta fecha horaria. Esto explicaría una producción tan alta para un día en apariencia con viento moderado, como indican las predicciones.

Modelo 11/03/2018-14:00:00

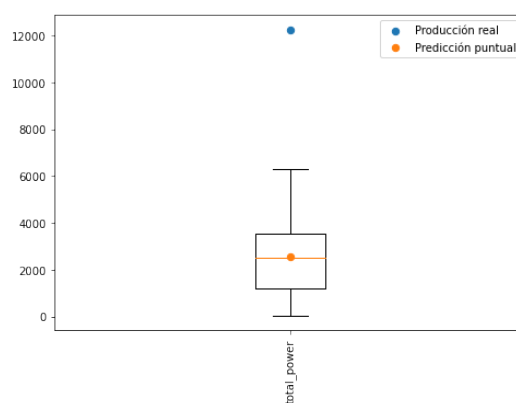
Este modelo predijo una potencia igual a 11582.38 kW, pero en Sotavento la potencia registrada fue 0. Veamos en la Figura 4.10 las estadísticas de las velocidades y las componentes del viento de los vecinos seleccionados.



(a) Vel10 y Vel100

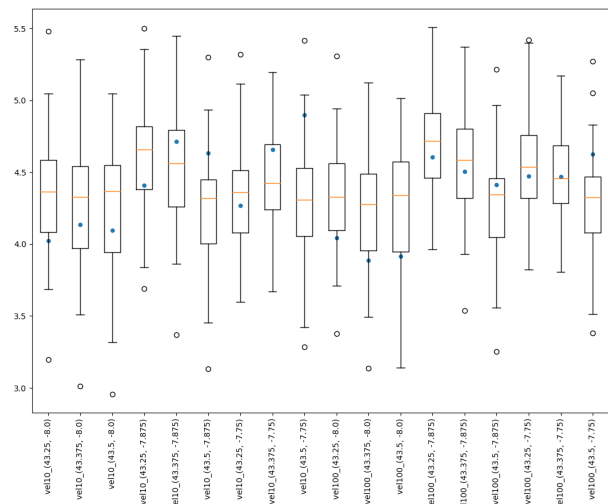


(b) 2t

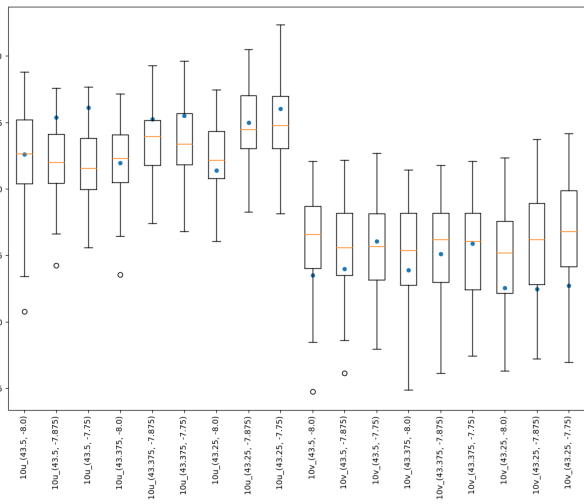


(c) Potecia

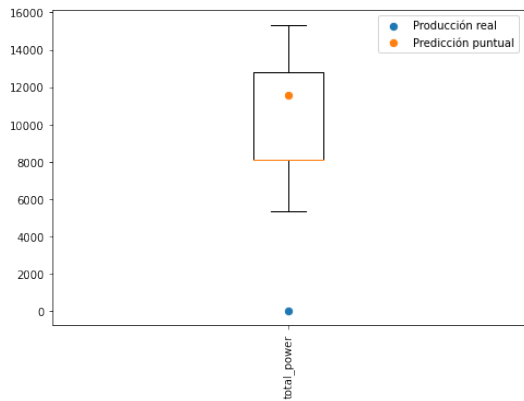
Figura 4.9: Estadísticas de la velocidad 4.9(a), la temperatura 4.9(b) y las potencias registradas 4.10(c) para el modelo 13/02/2018-20:00:00.



(a) Vel10 y Vel100



(b) 10u y 10v



(c) Potencia

Figura 4.10: Estadísticas de la velocidad a 10 y a 100 metros 4.10(a), las componentes u y v a 10 metros 4.10(b) y las potencias registradas 4.10(c) para el modelo 11/03/2018-14:00:00.

Todos las velocidades son altas y el dato a predecir se encuentra dentro del rango intercuartílico. Lo mismo ocurre con las componentes u y v del viento a 10 metros. Para las potencias que registran los vecinos seleccionados, vemos que toman la mayoría de los valores entre 8000 y 12000 kW, y, aunque la predicción no se acerca a la mediana, pertenece a su rango intercuartílico.

Con esas predicciones meteorológicas y basándonos en los puntos seleccionados, los aerogeneradores deberían producir energía. Además, si observamos las potencias registradas para las horas anteriores y posteriores a las 14:00:00 en la Figura 4.11, vemos que existen un total de 21 horas con potencias 0 registradas. Por tanto, todo parece indicar que se trata de una hora en la que las máquinas pararon su funcionamiento y esta particularidad explicaría un número notable de los errores del modelado local.

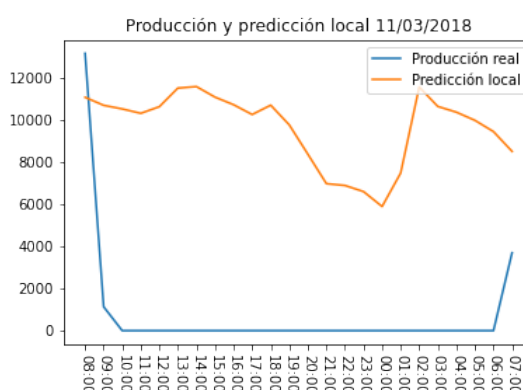


Figura 4.11: Potencias registradas horas antes y después de las 14:00:00h del día 11/03/2018.

4.4. Experimento 2

En este experimento se pretende evaluar el rendimiento de SVR utilizando el modelo local propuesto en el trabajo. Preprocesados los datos, se realizará una etapa de validación para obtener los valores de los hiperparámetros que mejor ajustan SVR sobre todos los datos de entrenamiento y se evaluará el modelo local seleccionando distinto número K de vecinos más cercanos. Finalmente, se evaluará el modelo local validando cada modelo local individual y se compararán los resultados de ambas evaluaciones.

4.4.1. Validación

Se lanza un proceso de validación cruzada de 5 particiones utilizando una búsqueda en rejilla sobre los hiperparámetros de SVR. La Tabla 4.3 muestra los conjuntos de hiperparámetros utilizados para la validación global. En la tabla, d es el número de dimensiones de los datos y σ es la desviación estándar de la función objetivo.

C	ϵ	γ
$\{10^k : 3 \leq k \leq 5\}$	$\{\sigma/2^k : 2 \leq k \leq 6\}$	$\{4^k/d : -2 \leq k \leq 1\}$

Tabla 4.3: Hiperparámetros de SVR y rejilla utilizada para encontrarlos.

Se han seleccionado estos valores conforme a [13], acotando por arriba y por abajo el rango que presentaban para reducir el tiempo computacional.

Al validar el conjunto de entrenamiento total, el menor error obtenido corresponde a trabajar con $C = 10^4$, $\epsilon = \sigma/2^6$ y $\gamma = 4^{-1}/d$.

Para la validación de cada modelo local individual del último experimento se ha utilizado la misma rejilla definida en la Tabla 4.3.

4.4.2. Evaluación

Al evaluar SVR sobre los datos de prueba de 2018, utilizando los valores de los parámetros descritos arriba, obtenemos un NMAE igual **6.35 %**.

Por su parte, la evaluación del modelo local seleccionando distintos K vecinos más cercanos de los 17544 datos totales de entrenamiento y aplicando el mismo método de porcentajes del experimento anterior, se presenta en la Tabla 4.4.

%	0.01	0.025	0.05	1	3	5	10	25	50
K	17	44	87	175	350	877	1754	4386	8772
NMAE (%)	6.79	6.64	6.60	6.53	6.46	6.37	6.36	6.35	6.34

Tabla 4.4: NMAE para SVR local seleccionando distintos número de vecinos.

4.4.3. Análisis de resultados

El modelo local propuesto mejora al aumentar el número de vecinos hasta estabilizarse a partir de $K = 877$, presentando un error similar al obtenido con SVR global. Debe tenerse en cuenta que el carácter local implícito de la SVR destaca los puntos vecinos de cada punto mediante el uso del núcleo RBF. Además, en la Figura 4.12 se observan algunas predicciones horarias obtenidas por el modelo local con 877 vecinos y por el modelo global. Apenas difieren las predicciones y ambas mantienen la misma tendencia.

Finalmente ilustraremos las predicciones horarias dadas por el modelo SVR global, el modelo SVR local con 350 vecinos y las producciones reales, para los mismos 4 días que en el experimento anterior.

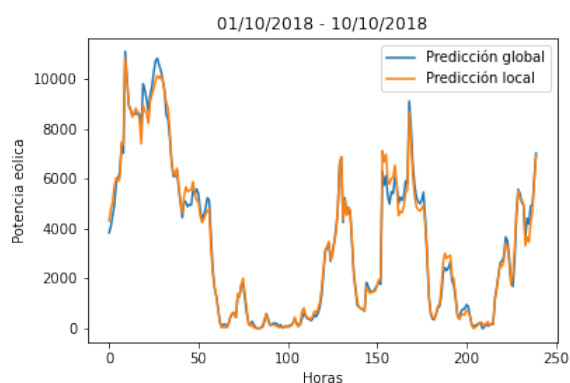
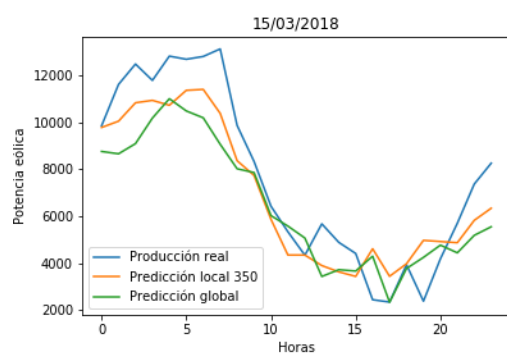
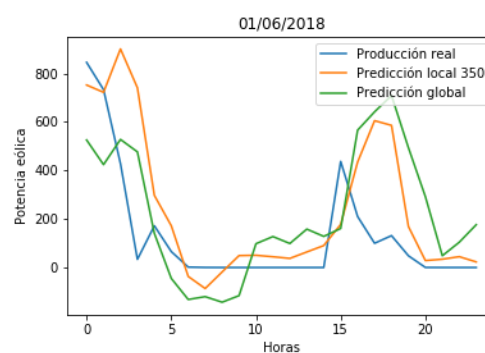


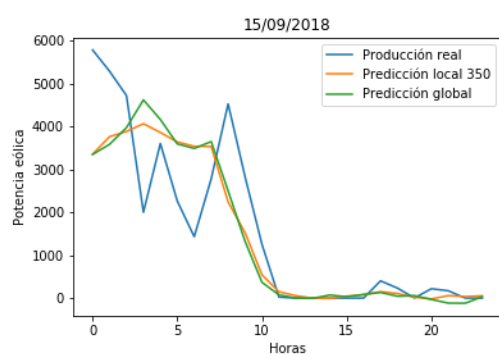
Figura 4.12: Predicciones horarias de los primeros 10 días de Octubre de 2018 obtenidas por el modelo SVR global y el modelo SVR local con $K=877$.



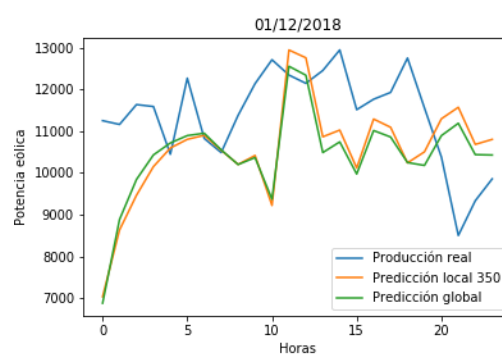
(a) 15 de Marzo de 2018



(b) 1 de Junio de 2018



(c) 15 de Septiembre de 2018



(d) 1 de Diciembre de 2018

Figura 4.13: Producciones y predicciones locales y globales con SVR en diferentes días a lo largo del año 2018.

Como podemos observar en la Figura 4.13, las predicciones del modelo local con 350 vecinos se mantienen cercanas a las predicciones del modelo global, y ambas se aproximan a las producciones reales siguiendo su tendencia.

Finalmente, se ha decidido evaluar el modelo de regresión local SVR validando cada modelo local individual con el conjunto de entrenamiento seleccionado por K -NN, para algunos K vecinos. La Tabla 4.5 muestra el error obtenido por cada modelo, obteniendo el menor error el modelo que selecciona $K=877$ vecinos, con un NMAE igual a **6.76 %**. Aunque no es un mal resultado, no se ha conseguido mejorar el modelo SVR global ni el modelo local, utilizando los hiperparámetros del modelo global. Parece que, siendo SVR muy sensible a los hiperparámetros, con tan pocos valores para ajustar el modelo no es capaz de generalizar bien. Por este motivo y por el gran coste computacional que ha supuesto el proceso de validación de cada modelo local individual, estos modelos no serán considerados como modelos competitivos en el trabajo.

%	1	3	5
K	175	350	877
NMAE (%)	7	7	6.76

Tabla 4.5: NMAE para SVR local validando cada modelo individual y variando K

4.5. Tiempo de cómputo

En el contexto de predicción de energía eólica, es necesario conocer con antelación suficiente la cantidad de energía que será generada. Con el objetivo de comparar los costes temporales, se ha calculado el tiempo que tarda cada modelo, ya sea Ridge Regression o SVR, en predecir 8760 potencias horarias. También se han medido los tiempos para realizar la validación de los dos modelos con los 17544 datos del conjunto de entrenamiento, así como para entrenar los modelos con los parámetros óptimos. Estos tiempos pueden observarse en la Tabla 4.6.

	Validación	Fit	Predict
RR Global	0.6 s	0.014 s	0.0009 s
SVR Global	~ 12h	24.52 s	8.46 s

Tabla 4.6: Tiempo de cómputo de RR y SVR global.

Por otra parte, se ha calculado el tiempo total en realizar los entrenamientos y las predicciones de los modelos locales correspondientes a los 8760 datos de prueba. Los resultados se muestran en la Tabla 4.7. El proceso de entrenamiento se ha realizado utilizando los parámetros óptimos del conjunto global sobre 44 vecinos para RR y 877 vecinos para SVR.

	Fit+Predict total	Fit+Predict medio
RR 44 vecinos	26 min	0.17 s
SVR 877 vecinos	36 min	0.24 s

Tabla 4.7: Tiempo de cómputo de RR y SVR local.

Observando estos resultados podemos empezar a sacar conclusiones.

La búsqueda de hiperparámetros en SVR global es muy costosa, con un coste temporal de 12h aproximadamente. RR, por su parte, al utilizar un modelo local con 44 vecinos tarda alrededor de 26 minutos en realizar la validación global y obtener las predicciones. Por tanto, Ridge Regression local supone una gran mejora en términos computacionales respecto a SVR global, a costa de un porcentaje de error de 5 décimas mayor. Sin embargo, si obviamos el proceso de validación, SVR global tarda únicamente 24,52 segundos frente a los 26 minutos de RR local con 44 vecinos.

Por otro lado, podemos comparar SVR local con 877 vecinos con RR local con 44. En este caso, la diferencia de coste computacional no es muy significativa si no tenemos en cuenta el proceso de validación. En cuyo caso, siempre convendría utilizar RR local con 44 vecinos.

Por último compararemos SVR global con el modelo SVR local. En este caso, el modelo global supone siempre una gran mejora en términos computacionales respecto al modelo local SVR tomando cualquier número de vecinos, así como un mejor resultado predictivo.

CONCLUSIONES Y TRABAJO FUTURO

El continuo desarrollo de las energías renovables, y en concreto, de la energía eólica, ha hecho necesario el uso de herramientas de predicción que permitan conocer con antelación la cantidad de energía eólica que será inyectada en la red. El aprendizaje automático es una herramienta muy útil en este contexto, permitiendo estimar la producción de energía futura utilizando como información de entrada predicciones meteorológicas en algunas coordenadas de la región de interés. Además, cuando se recurre al modelado local, los pronósticos de energía suelen ser más precisos ya que los modelos no están dominados por muestras correspondientes a módulos de viento bajos. En este trabajo, se ha aplicado un modelo de regresión local que consiste en definir un modelo diferente para cada punto del conjunto de prueba conteniendo la información más parecida al punto a predecir al seleccionar los K vecinos próximos del conjunto completo de entrenamiento. Para medir la similitud entre los vectores meteorológicos se ha utilizado la distancia euclídea. Este modelo de regresión local se ha realizado utilizando el algoritmo de regresión lineal RR y el algoritmo de regresión no lineal SVR mediante un núcleo RBF.

Para el modelo de regresión lineal Ridge Regression utilizando el parámetro de regularización $\lambda = 100$, el menor error obtenido es 6.83 % y corresponde a utilizar $K = 44$ vecinos del conjunto total de entrenamiento. De hecho, el error aumenta al aumentar el número de vecinos aproximándose al NMAE 8.44 % de RR para el modelo global. La matriz de correlación de las variables regresoras y la variable objetivo, marcó a la velocidad del viento como la variable más correlacionada con la potencia eólica. En todos los modelos la selección de vecinos próximos se realizó satisfactoriamente, y se observó la selección de vientos bajos para potencias nulas y vientos altos para potencias altas. Analizando los modelos que presentaban altos residuos, se consiguió explicar el error para aquellas horas en las que se realizaron paradas de mantenimiento o para las que el error provenía de las propias predicciones meteorológicas. Respecto al tiempo de cómputo necesario, el proceso de validación se llevó a cabo en únicamente 6 segundos y el proceso entrenamiento y evaluación supuso un coste de 26 minutos. Dado que la validación de RR es poco costosa, se concluye que este modelo de regresión local supone una gran ventaja puesto que permite actualizar los parámetros de manera frecuente, ya sea cada mes anterior o incluso cada día. Además, cabe resaltar que, gracias al modelo de regresión local propuesto, un modelo lineal como RR se convierte en un modelo competitivo en predicción de

energía eólica.

Para el modelo de regresión no lineal SVR utilizando los hiperparámetros $C = 10^4$, $\epsilon = \sigma/2^6$ y $\gamma = 4^{-1}/d$, al contrario que con RR, aumentar el número de vecinos próximos utilizados en el entrenamiento supone una mejora del NMAE desde 6.79 % con 17 vecinos hasta estabilizarse a 6.37 % con $K = 877$. El NMAE obtenido para el modelo global SVR con estos mismos parámetros es igual a 6.35 %. Observando esta similitud del NMAE y argumentando en base al carácter local implícito de la SVR, se puede concluir que son aproximadamente 877 los datos significativos para la estimación considerados en el modelo global con el núcleo gaussiano. Respecto al coste de cómputo, se necesitaron aproximadamente 12 horas para el proceso de validación y por tanto, SVR requiere de una mayor planificación a la hora de actualizar los parámetros. Por su parte, el proceso de entrenamiento y evaluación del modelo de regresión local con $K = 877$ se llevó a cabo en 36 minutos, mientras que este mismo proceso para el modelo global sólo supuso un coste de 33 segundos. Esta diferencia de tiempo no es notable teniendo en cuenta que, en aplicaciones reales de predicciones de energía eólica a medio plazo, sólo se requeriría entrenar el modelo con las previsiones meteorológicas previstas para las horas futuras. Por tanto, el modelo de regresión local no lineal SVR iguala al modelo local pero no consigue mejorarlo.

Como posibles mejoras en el modelo, se podría incluir el número K de vecinos seleccionados para entrenar cada modelo local individual en el propio proceso de validación. Además, una futura línea de investigación es utilizar previsiones meteorológicas dadas por modelos ensembles NWP para dar estimaciones de la incertidumbre asociada a las predicciones de energía. De esta manera, se podrían comparar los modelos de regresión local propuestos con los modelos globales de acuerdo a la incertidumbre asociada a cada uno de ellos. También podría utilizarse otra medida de distancia para la selección de K -NN, e incluso estudiar cómo se comporta este método para predicciones de otras energías renovables, como la solar.

BIBLIOGRAFÍA

- [1] "Global wind report 2019." <https://gwec.net/global-wind-report-2019/>. Accedido 09-03-2021.
- [2] "Asociación empresarial eólica." <https://www.aeeolica.org/sobre-la-eolica/la-eolica-espana>. Accedido 09-03-2021.
- [3] "Instituto para la diversificación y ahorro de la energía." <https://www.idae.es/informacion-y-publicaciones/plan-nacional-integrado-de-energia-y-clima-pniec-2021-2030>. Accedido 09-03-2021.
- [4] X. Wang, P. Guo, and X. Huang, "A review of wind power forecasting models," *Energy Procedia*, vol. 12, pp. 770–778, 2011. The Proceedings of International Conference on Smart Grid and Clean Energy Technologies (ICSGCE 2011).
- [5] J. Jung and R. P. Broadwater, "Current status and future advances for wind speed and power forecasting," *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 762–777, 2014.
- [6] C. Gallego, P. Pinson, H. Madsen, A. Costa, and A. Cuerva, "Influence of local wind speed and direction on wind power dynamics – application to offshore very short-term forecasting," *Applied Energy*, vol. 88, no. 11, pp. 4087–4096, 2011.
- [7] M. De Felice, A. Alessandri, and P. M. Ruti, "Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models," *Electric Power Systems Research*, vol. 104, pp. 71–79, 2013.
- [8] M. Duran, D. Cros, and J. Santos, "Short-term wind power forecast based on ARX models," *Journal of Energy Engineering-ASCE*, vol. 133, 09 2007.
- [9] S. Jung and S.-D. Kwon, "Weighted error functions in artificial neural networks for improved wind energy potential estimation," *Applied Energy*, vol. 111, pp. 778–790, 2013.
- [10] J. Zhang, J. Yan, D. Infield, Y. Liu, and F. Sang Lien, "Short-term forecasting and uncertainty analysis of wind turbine power based on long short-term memory network and Gaussian mixture model," *Applied Energy*, vol. 241, pp. 229–244, 2019.
- [11] J. Wang, W. Yang, P. Du, and T. Niu, "A novel hybrid forecasting system of wind speed based on a newly developed multi-objective sine cosine algorithm," *Energy Conversion and Management*, vol. 163, pp. 134–150, 2018.
- [12] Y.-Y. Hong and C. L. P. P. Rioflorida, "A hybrid deep learning-based neural network for 24-h ahead wind power forecasting," *Applied Energy*, vol. 250, pp. 530–539, 2019.
- [13] C. Ruiz, C. M. Alaíz, and J. R. Dorronsoro, "Multitask support vector regression for solar and wind energy prediction," *Energies*, vol. 13, no. 23, 2020.
- [14] S.-F. Wu and S.-J. Lee, "Employing local modeling in machine learning based methods for time-series prediction," *Expert Systems with Applications*, vol. 42, no. 1, pp. 341–354, 2015.
- [15] C. Alaíz, A. Barbero, Fernández Pascual, and J. Dorronsoro, "High wind and energy specific models for global production forecast," 01 2009.
- [16] "Ensemble prediction system." <https://www.ecmwf.int/sites/default/files/elibrary/2012/14557-ecmwf-ensemble-prediction-system.pdf>. Accedido 10-03-2021.
- [17] "Física del caos en la predicción meteorológica." http://www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/publicaciones/Fisica_del_caos_en_la_prediccion_meteo/19_El_Centro_Europeo_de_Prediccion_a_Medio_Plazo.pdf. Accedido 10-03-2021.
- [18] "Centro europeo de previsiones meteorológicas a medio plazo." <https://www.ecmwf.int/>. Accedido 10-03-2021.
- [19] H. Shaker, H. Zareipour, and D. Wood, "On error measures in wind forecasting evaluations," in *2013 26th IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pp. 1–6, 2013.
- [20] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.
- [21] A. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 08 2004.

- [22] W. S. Cleveland and S. J. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 596–610, 1988.

APÉNDICES

ESTADÍSTICAS DESCRIPTIVAS

Se van a mostrar las estadísticas de las variables que no fueron ilustradas en el trabajo para los 4 modelos comentados en el Capítulo 5. Finalmente, la Sección A.5 visualiza las 12 estadísticas más significativas de un modelo que registra potencias altas y una buena predicción para compararlo con el modelo con fecha horaria 13/02/2018-20:00:00 descrito en la Sección 4.3.

A.1. Estadísticas modelo 31/12/2018 - 20:00:00

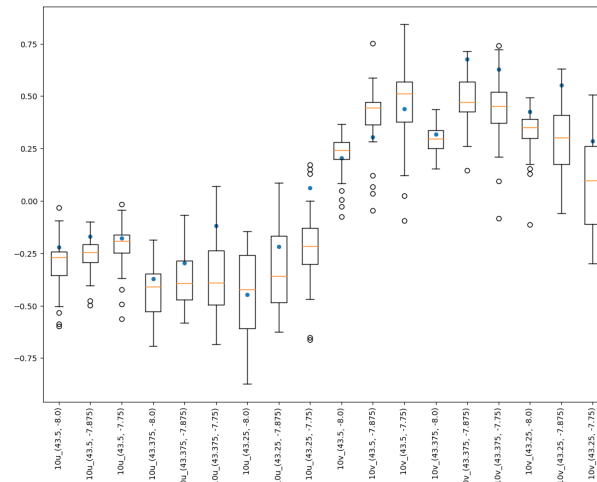
En la Figura A.1 se ilustran las estadísticas de las componentes u y v a 10 metros, de las temperaturas a 2 metros y la presión a nivel de superficie en las 9 coordenadas que rodean Sotavento para el modelo con fecha horaria 31/12/2018 - 20:00:00.

A.2. Estadísticas modelo 09/01/2018 - 07:00:00

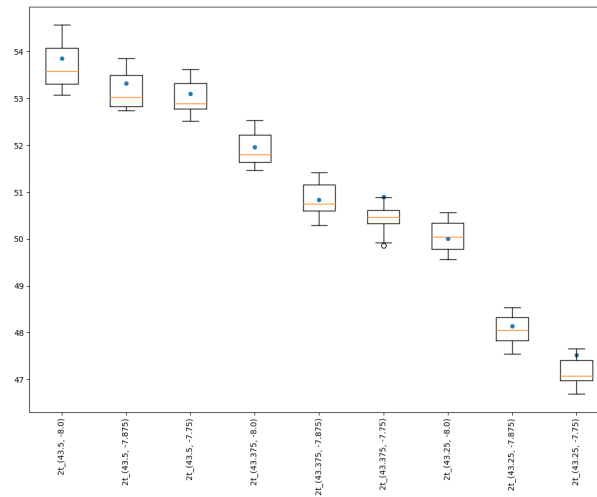
En la Figura A.2 se ilustran las estadísticas de las temperaturas a 2 metros y la presión a nivel de superficie en las 9 coordenadas que rodean Sotavento para el modelo con fecha horaria 09/01/2018 - 07:00:00.

A.3. Estadísticas modelo 13/02/2018 - 20:00:00

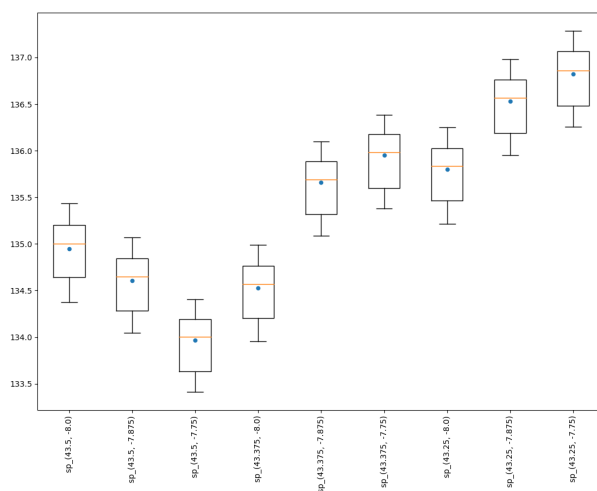
En la Figura A.3 se ilustran las estadísticas correspondientes a las componentes del viento u y v a 10 y 100 metros, así como la presión en las 9 coordenadas que rodean Sotavento para el modelo con fecha horaria 13/02/2018 - 20:00:00.



(a) 10u y 10v

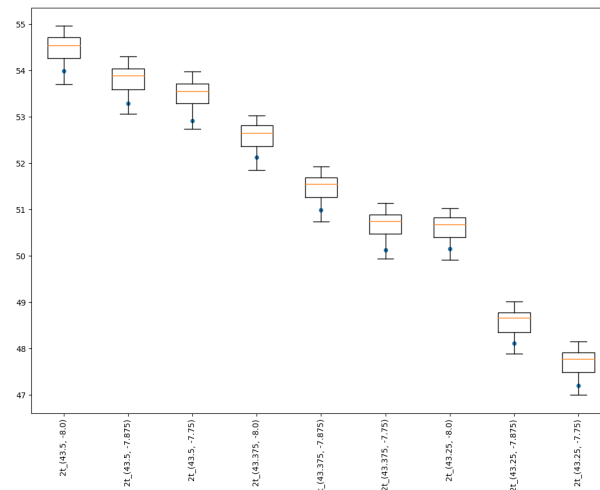


(b) 2t

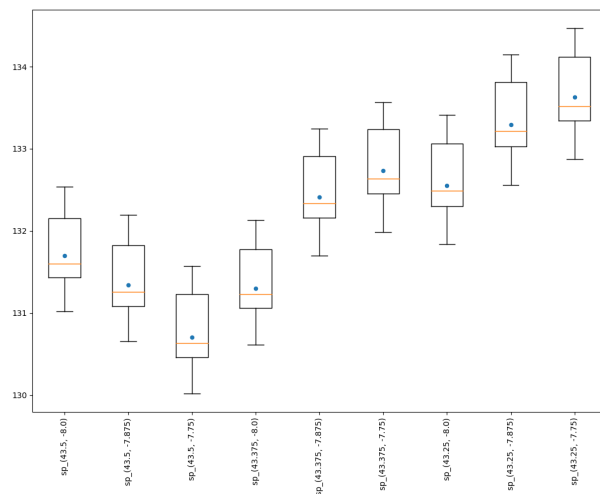


(c) sp

Figura A.1: Estadísticas de las componentes u y v a 10 metros A.1(a), de las temperaturas a 2 metros A.1(b) y la presión a nivel de superficie A.1(c) para el modelo 31/12/2018 - 20:00:00.

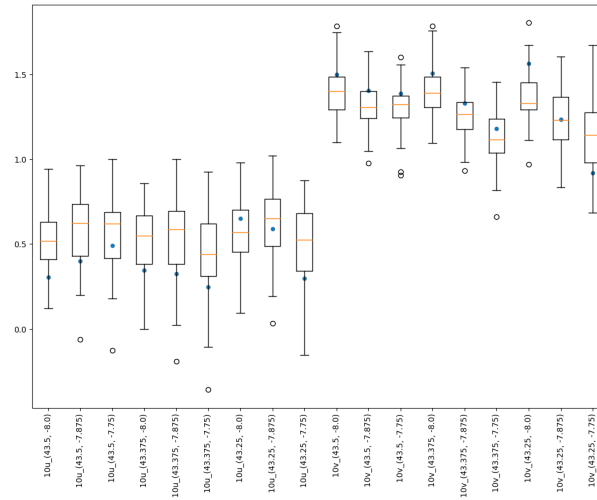


(a) 2t

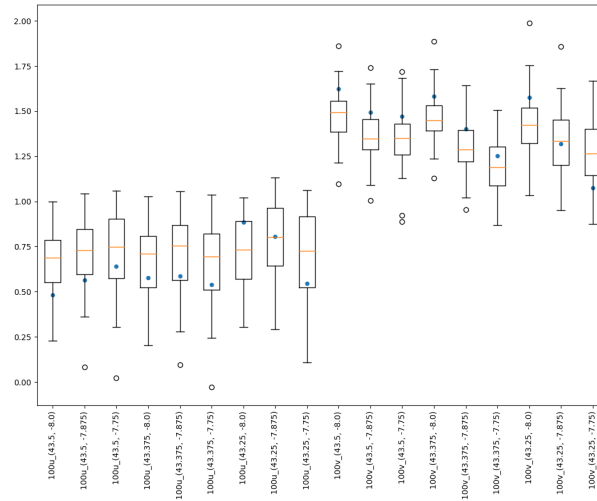


(b) sp

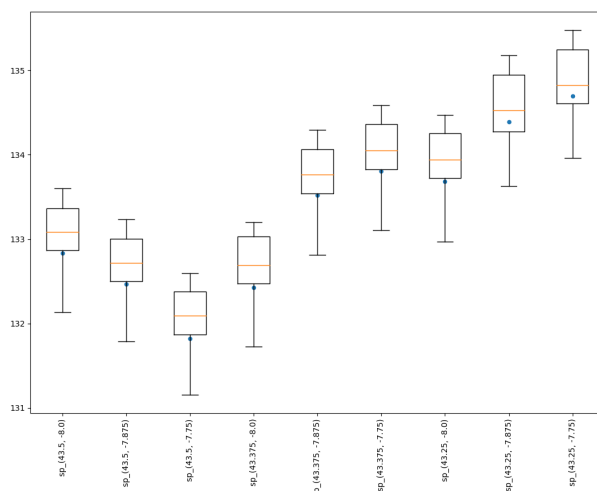
Figura A.2: Estadísticas de las temperaturas a 2 metros A.2(a) y la presión a nivel de superficie A.2(b) para el modelo 09/01/2018 - 07:00:00.



(a) 10u y 10v



(b) 100u y 100v



(c) sp

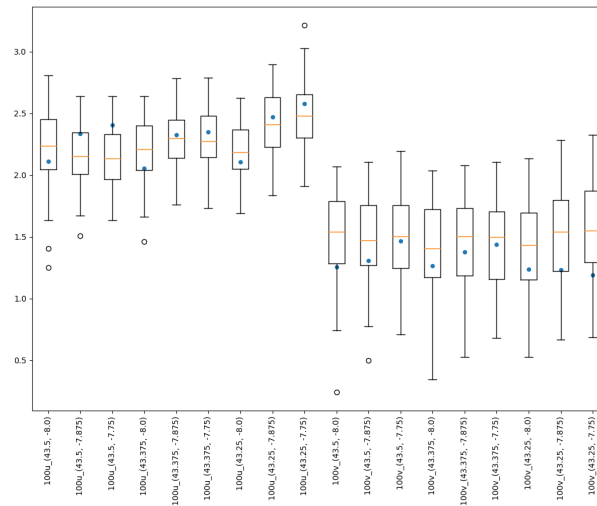
Figura A.3: Estadísticas de las componentes u y v a 10 metros A.3(a) y a 100 metros A.3(b) y de la presión a nivel de superficie A.3(c) para el modelo con fecha horaria 13/02/2018 - 20:00:00.

A.4. Estadísticas modelo 11/03/2018 - 14:00:00

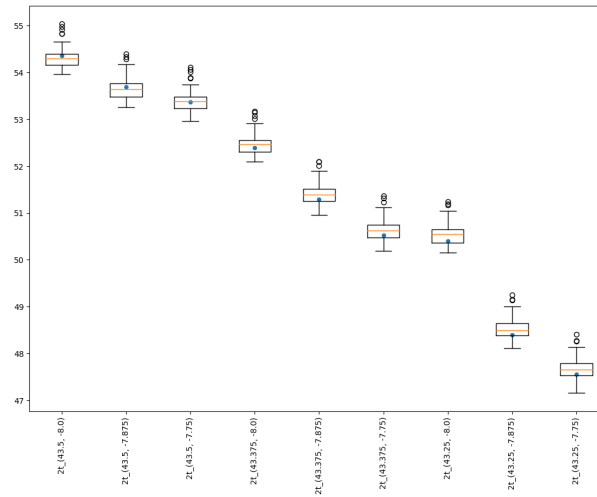
En la Figura A.4 se ilustran las estadísticas correspondientes a las componentes del viento u y v a 100 metros, así como la temperatura y la presión en las 9 coordenadas que rodean Sotavento para el modelo con fecha horaria 11/03/2018 - 14:00:00.

A.5. Estadísticas modelo 15/12/2018 - 07:00:00

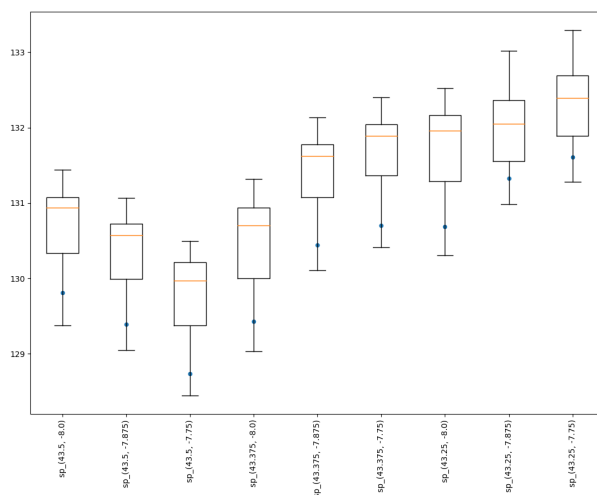
En la Figura A.5 se ilustran las estadísticas correspondientes a las 12 variables más significativas presentes en la matriz de correlación. También se muestran las estadísticas de las variables objetivos correspondientes.



(a) 100u y 100v

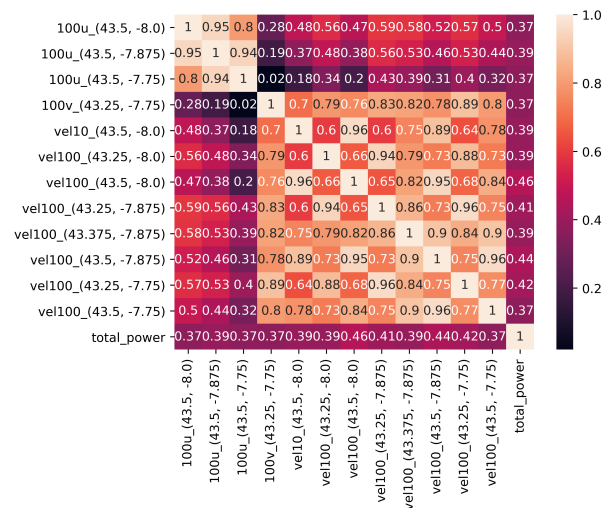


(b) 2t

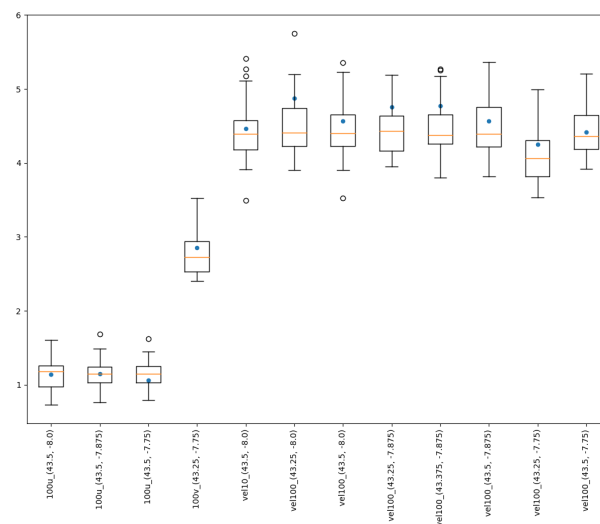


(c) sp

Figura A.4: Estadísticas de las componentes u y v a 10 metros A.4(a) y a 100 metros A.4(b) y de la presión a nivel de superficie A.4(c) para el modelo con fecha horaria 11/03/2018 - 14:00:00.



(a) Matriz correlación



(b) 12 variables significativas



(c) Potencia

Figura A.5: Estadísticas de las 12 variables más significativas A.5(b) presentes en la matriz de correlación A.5(a) y estadísticas de la variable objetivo A.5(c) para el modelo 15/12/2018 - 07:00:00.

